1

The Puzzle of Seeing

1.1 Why Do Things Look the Way They Do?

Why do things appear to us as they do? We don't even have a clear idea of what kind of story would count as an answer. The whole notion of "appearing" seems problematic. On one hand, it is obvious that things appear to us as they do because, barring illusions, *that's the way they are*! On the other hand, it is also clear that a particular thing's being as it is, is neither necessary nor sufficient for our *seeing* it as we do. We know that things can look quite different to us in different circumstances, and perhaps they do look different to others. So it is not unreasonable for us to ask what is responsible for our seeing things as we do, as opposed to seeing them in some other way.

Despite the dramatic progress that has been made in the study of visual perception in the past half century, the question of why we see things as we do in large measure still eludes us. The question of what and how and why we see are daunting. Surely, the pattern of light arriving at our eyes is responsible for our visual perception. Must this be so—is light both necessary and sufficient for perception? Could we not also "see" if our eye or our brain were electrically stimulated in the right way? And what of the experience of seeing: Is that constitutive of vision; is that what visual perception *is*? Would it make any sense to ask what is the product or even the purpose of visual perception? Could there be full-blooded visual perception in the absence of any awareness of something being seen, without a visual experience? The mystery of the experience of seeing is deep and is at the heart of our understanding (or failure to understand) the nature of consciousness itself. Is it possible to have a scientific understanding of vision without first understanding the mystery of consciousness? The scientific world thinks it is, and it has already made a great deal of progress in acquiring such an understanding. But is this because it has presupposed a view of what it is to see—a set of tacit assumptions about such things as the relation between our experience of seeing and the nature of the information processing performed by the visual system?

I do not intend this book to be about consciousness, or even about our conscious visual experience, because I believe there is little that science can say about this notion at the present time. That's not to say that it is not of the greatest importance and perhaps even central to understanding human nature. It is also not to say that there is nothing worthwhile to be said on the topic of consciousness, since consciousness has become a very active topic of scholarship and a great deal is being said, much of it quite fascinating. Nonetheless, most of what is being said is by way of preliminary scene setting and conceptual clarification. It's about such surprising empirical findings as those showing that certain functions can be carried out without conscious awareness. A lot of the discussion is also about what consciousness is not. It's about such things as why a theory simply misses the point if it says that consciousness *is* such and such a brain property (a certain frequency of brain waves or activity in a certain location in the cortex) or a particular functional property (such as the contents of short-term working memory or the mind's observation of its own functioning). The only part of this discussion that will concern us will be how the content of our experience when we see, visualize, and think misleads us and contaminates many of our scientific theories of vision and of related processes such as visualizing and imagining. For this reason I devote much of the present chapter to a discussion of what vision provides to the mind. The closely related question of how the cognizing mind affects visual perception is raised in chapter 2, and some of that discussion takes us back to the troublesome notion of the nature of visual experience.

1.2 What Is Seeing?

One reason why understanding vision is so difficult is that we who are attempting to understand the process are so deeply embedded in the phenomenology of perception: We know what it *feels* like to see. We look out and see the world, and we cannot escape the impression that what we have in our heads is a detailed, stable, extended, and veridical display that corresponds to the scene before us. Of course, most of us have also seen enough examples of so-called "optical illusions," so we are prepared to admit that what we see is not always what is truly the case. Yet at the same time we have much more difficulty shedding the view that in our heads is a display that our inner first-person self, or our cognitive homunculus, observes. There are other phenomena relating to our experience of seeing a "picture in our head" that are even more problematic. These include the similar experience that we have without any visual input: the experience that accompanies mental imagery or visual thinking. The more we analyze what must be going on and the more we examine the empirical evidence, the more puzzling the process becomes and the less tenable our intuitions. Indeed, we find not only that we must dispense with the "picture in the head," but that we must also revise our ideas concerning the nature of the mechanisms involved in vision and concerning the nature of the internal informational states corresponding to percepts or images. What can never serve as a theory of vision is a theory that says that vision creates a copy of the world inside the head, as the Kliban cartoon in figure 1.1 suggests is the case with a cat. The understanding that this sort of theory will not do is what makes this cartoon funny. Yet it is nontrivial to say what exactly is wrong with a theory that even remotely resembles this sort of story. This I will attempt in the present book, mostly in this chapter and in chapters 6 and 7.

In what follows I will examine some of these counterintuitive aspects of the process of visual perception and mental imagery. For now the following examples will suffice to warn us that our intuitions are a notoriously bad source of ideas as to how the visual system works. The message of these examples is that we should not be surprised to find that our scientific theories will look quite different from how we might imagine them when we try to be faithful to how vision seems to us from the inside—to the phenomenology of visual perception.

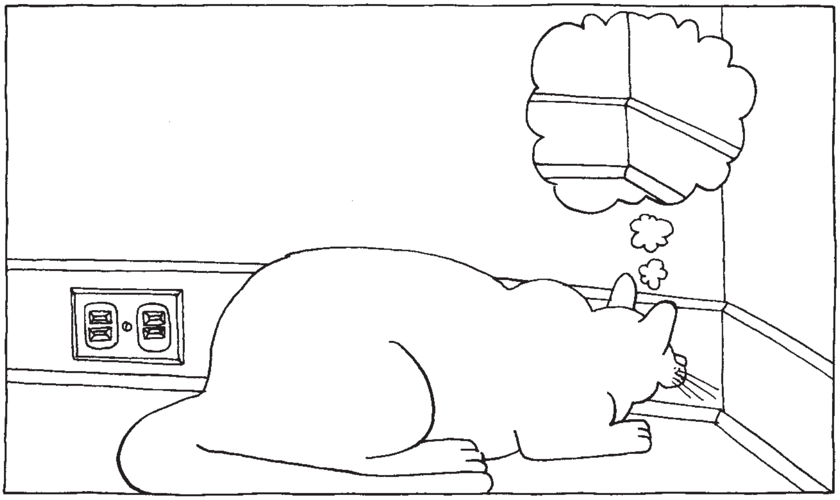


Figure 1.1
A theory of vision such as this is a nonstarter, even for a cat! B. Kliban (American, 1935–1990). From the book *Cat*, by B. Kliban. Used by permission. All rights reserved. Copyright by Judith K. Kliban.

1.3 Does Vision Create a "Picture" in the Head?

1.3.1 The richness of visual appearances and the poverty of visual information

Let's call our conscious experience of how things seem to us when we look at them, the "phenomenal" content of our perception. As we look around, the phenomenal content of our perception is that of a detailed and relatively stable panorama of objects and shapes laid out in three dimensions. Even without turning around, we experience a broad expanse (about 180 degrees of panorama), full of details of the scene: its colors and textures, its shapes and boundaries, and the meaningful things that populate our visual scene—the familiar objects and people that we instantly recognize. Even if there were little or nothing in the scene that we recognized as familiar, say if we had just landed on the surface of Mars, we would still have no trouble seeing shapes and surfaces. We would see a variety of individual objects, set against some background that remained perceptually secondary (i.e., we would experience what Gestalt psychologists call a "figure-ground" separation). We would see each of these objects as having a certain shape and consisting of parts arranged in some spatial relation to one another. We would see some of the objects as further away and some as closer, with the closer objects partially occluding our view of the further objects. We would see that the partly occluded objects continued behind the closer ones; we would *not* see the occluded objects as partial objects or as having the shape of the visible fragment, though it is physically possible that this could in fact be their shape. The phenomenal content of our perception would continue to be that of a world of three-dimensional objects, even though most of every object would be hidden from our view, either by other objects or by the front of the object itself. If we could turn freely to inspect the scene around us, there would be no sharp discontinuity between the part of the scene currently on our retina and the entire 360 degrees of the layout (e.g., we could accurately point to objects behind us, as Attneave and Farrar, 1977, showed).

This phenomenal experience is, as far as we know, universal to our species and probably innate. We don't give it a second thought, because it seems to us that we are seeing what there is to see. But even a cursory examination makes it abundantly clear that much more is going on than we might assume (I am tempted to say that there is more to vision than meets the eye). Consider what the brain has to work with in achieving this familiar experience. The light-sensitive surfaces of the eye (the retinas) are two-dimensional, so the sense of depth must come from other sources of information. We know that at least part of the information comes from the difference between the patterns that the two eyes receive, but why (and how) does this produce the experience of seeing a three-dimensional world? No matter how well we understand the mechanism of stereo perception (and it is one of the most studied problems in visual science), we are very far from breaking through the mystery of this question. The story gets even murkier as we further examine the information that the brain receives from the eyes. The retinas themselves are not uniform. Only a small central region (the fovea), about the size of the area covered by your thumb held at arm's length, has sufficient acuity to recognize printed characters at the normal reading distance. Outside of that region our visual acuity drops off rapidly, and by the time we get to where the edge of a movie screen would normally fall in our field of vision, acuity is so poor that if we thus saw the world generally, we would be considered legally blind. As we move off from the central fovea, the eye also becomes color blind, so almost all color information comes from the tiny area of the fovea (and what color reception there is varies in its degree of responsiveness to the yellow-green dimension depending on how far out from the fovea it is). Moreover, our eye's focal length differs considerably for red and blue colors, so one end of the spectrum is invariably out of focus by about the degree of magnification of off-the-shelf reading glasses. There is also a region of the retina, considerably larger than the fovea and lying about 10 to 13 degrees away, where the retinal nerve fibers come together to form a cable to the brain. This region has no receptors: it is our blind spot. It is easy to show that no information is registered at the location of the blind spot (look at figures 1.2 and 1.3), yet we are unaware of the blind spot: it does not interfere with our

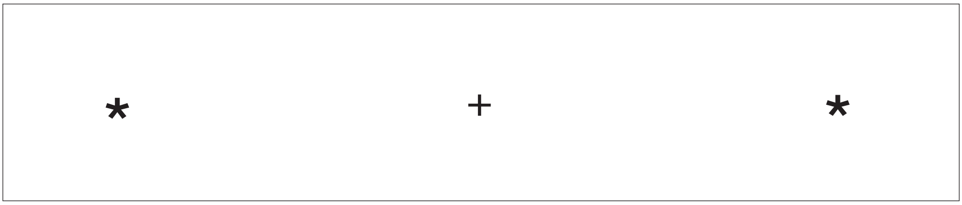


Figure 1.2
If you close your right eye and look at the plus sign with your left eye at a distance of about 10 to 12 inches from the paper (varying the distance as you experiment) you will find that the asterisk on the left disappears from view at some appropriate distance. If you repeat this with your right eye the asterisk on the right will disappear. This is because they fall on the blind spot of each eye. Now repeat the experiment on figure 1.3.

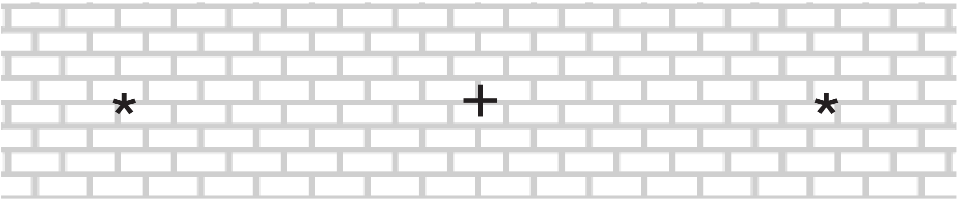


Figure 1.3
If you repeat the experiment from figure 1.2 on this figure, held at the same distance where the asterisk disappeared on the previous figure, you may find that the bricklike pattern, though somewhat indistinct, remains visible without a *gap while the asterisk disappears*. This is a case of what has been called "filling in." But is there some place in your mind/brain where there is an inner display that has filled in the missing pattern?

phenomenal experience of a uniform perceptual world. Even when the spot is located over an elaborate pattern, we do not see a hole in the pattern. In fact, we even see objects move through the blind spot without discontinuity, and we can locate the moving object precisely as being inside the blind spot at the appropriate time (Cai and Cavanagh, 2002). This phenomenon, which many people refer to as "filling in," provides some important clues as to how vision works and how vision, *as an information-processing system*, relates to our phenomenal experience of seeing. We will return to this most perplexing question later.

The properties of the retina (including its blind spot and other distortions of incoming information because it is not flat and the distribution of rods and cones is not uniform) already provide some reason to be concerned about how the brain gets to have such a large expanse of visual experience.[1] But it gets much worse. The eyes are in constant motion, jumping around in rapid saccades several times in each second and generally spending only a fraction of a second gazing in any one direction. The retina, our primary contact with the visual world, is continually being smeared with moving information (moving so rapidly that the nervous system cannot assimilate any detailed information during the rapid saccade, which can take as little as 3 hundredths of a second to sweep its path). And yet the world does not appear to move or flicker, and indeed, we are typically unaware of the saccades. How do we see a rich and stable visual panorama in the face of such dynamic and impoverished information?

The intuitive answer that almost universally leaps to mind is that although the *retina* may get a degraded, nonuniform, rapidly changing,

[1] In 1604 when Johannes Kepler first described how an image is formed on the retina (Lindberg, 1976) people began to worry about how we manage to see the world right-side-up when the image on the retina is upside-down. These days this no longer bothers most people because they implicitly understand that what counts as *up* for us is determined by how the brain *interprets* the retinal image and how it coordinates properties of the image with our actions on the world. This downplaying of physical image-properties in relation to both our phenomenal experience and our motor behavior toward the perceived world marks the beginning of an appreciation that information processing and phenomenal experience are a long way from the retina and that much goes on in the interval. Yet, as we will see presently, the temptation to mistake the retinal image for the percept continues even today.

peephole view of the world, you, the one who does the seeing, do not receive such impoverished information. What you see is a uniformly detailed, gapless, panoramic, stable view of the world—rather like a three-dimensional picture—built up from the sketchy unstable inputs from the two eyes. This is the first-person view, the world that your "self" (the subject in your claim "I see") gets to examine and enjoy. It is impossible for us to view what happens in our own perception in any other way. I think, my *eyes* may be moving, and they may have poor resolution with a blind spot and all that, but *I, the person observing these events, do not have any of these problems*. What I see is a 3D layout of surfaces and objects that have colors and shapes, and the entire scene stands still and covers a wide panorama (180 or more degrees). Consequently, so the argument goes, there must be something that has these properties of breadth, depth, and stability, and where else could it be but in the head? Enter what Dan Dennett picturesquely calls the "Cartesian Theater," after René Descartes, who, by implication (though not explicitly), proposed such an inner image or screen onto which the eye projects its moving peephole view and paints the larger picture.

But, tempting as it is, the Cartesian Theater creates far more problems than it solves. Dennett (1991) discusses a number of difficulties raised by this "inner eye" idea and shows that it leads to one conceptual impasse after another. The whole idea of an inner screen rests on a well-known fallacy, called the "intentional fallacy" in philosophy and sometimes the "stimulus error" in the structuralist psychology of Wundt and Titchener (in the case of Wundt and Titchener, "stimulus error" meant attributing to one's introspective experience the properties one knows the objective stimulus to possess). The temptation to make the mistake of attributing to a mental representation the properties of what it represents is difficult to avoid. This issue arises in an extreme form in discussions of mental imagery where the temptation appears to be very nearly inescapable (I will return to it in chapters 6 to 8; for an extensive discussion of this point, see Pylyshyn, 2002). What I propose to do in the rest of this chapter is to show that even if it does not create conceptual-philosophical problems, the inner-screen notion is at odds with some well-established facts about human vision. In the course of this discussion I will present a number of experimental results that will serve us later when I will return to the Cartesian Theater in connection with the idea of a mental image, an idea, as it is understood by many contemporary thinkers, that relies heavily on the assumption that there is a Cartesian Theater with a screen and a projector and a homunculus or "mind's eye" sitting in the audience.

While a misinterpretation of our phenomenal experience may be what drives us to the assumption of an inner display in the first place, it is not the only consideration that keeps many psychologists committed to it. In a great many cases the content of phenomenal experience also has observable consequences in objective measures of visual processing. In fact, phenomenal experience plays a central role in the methodology of visual science insofar as theories of vision are typically concerned with explaining the nature of our phenomenal experience. This in itself raises problems that will occupy some of our attention later. For now let us stay with the question of why many scholars of visual perception tacitly assume an inner display in attempting to understand how vision works.

1.3.2 Some reasons for thinking there may be an inner display

The overriding reason for believing in an inner display or image or Cartesian Theater is that the information on the retinas is so totally discrepant from the phenomenal experience of perception. We have already alluded to the peephole scope of retinal information, its rapidly changing contents, and its unnoticed blind spot that gets filled in for phenomenal experience. Then there are frequently noted completion phenomena, where familiar forms appear to get filled in when parts of them are occluded (as in figure 1.4), or where even unfamiliar forms appear to be filled in with illusory contours (illustrated in figure 1.5), or where there is so-called amodal completion (figure 2.5), which will be discussed later. This filling in is a subjective impression in the case of the blind spot, since there is no functional information available for the particular part of the scene corresponding to the scotoma. But in other cases it's not so obvious that *no* information is involved, even though there may be no local information at a particular site. For example, in the case of partially occluded figures, such as in figure 1.5, it is possible that the mind provides the missing information and actually restores the image, if not on the retina, then at some subsequent locus in the brain. In figure 1.4 the missing parts of the words don't just seem to be there, they are functionally present insofar as we are actually able to recognize and read the words.

So-called illusory or virtual contours (such as those seen in the figure on the right of figure 1.5) not only have a phenomenal existence; they

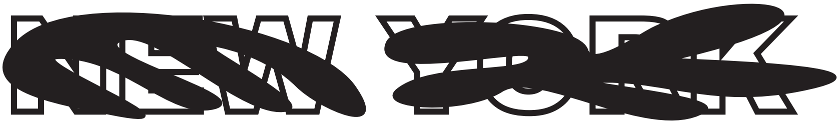


Figure 1.4
Despite the large amount of missing information, the familiar words are easily discerned. Are they “restored” by the visual system on some inner display?

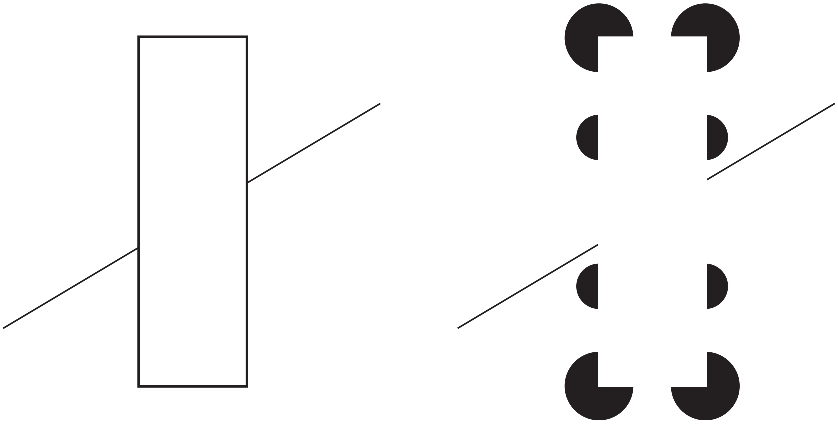
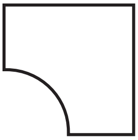


Figure 1.5
The Pogendorff illusion works as well with virtual (illusory) lines as with real ones. The oblique lines in both figures do not looked aligned, even though they are.

act in many ways as though they were actually present in the figures. Take, for example, the Pogendorff illusion, in which an oblique line that crosses a column appears to be broken and not aligned with its geometrical continuation (figure 1.5). When subjects are asked to adjust one part of the diagonal line so it appears to be continuous with the other part, they tend to set the lower line systematically higher than it should be for geometrically correct continuation. This phenomenon happens equally when the column is made up of virtual or illusory lines, as in figure 1.5.

Similarly, one can see that the “completed” figure is the one that is visually prominent by considering the problem of finding a given figure in a jumble (a kind of Where’s Waldo game). Consider the simple figure shown here:



Can you find that target in the jumble in figure 1.6? You may find it easy to identify one such instance, despite the fact that part of it is actually cut off by a circle. But there is another one that is harder to find because the visual system has "completed" it as a square, by adding the part that is hidden by the circular disk.

There are also many examples where visual properties are interrelated or *coupled* (to use Irvin Rock's term). Such visual "couplings" may depend on aspects of the perception that do not exist objectively on the retina. For example, the virtual rectangle created by the array of

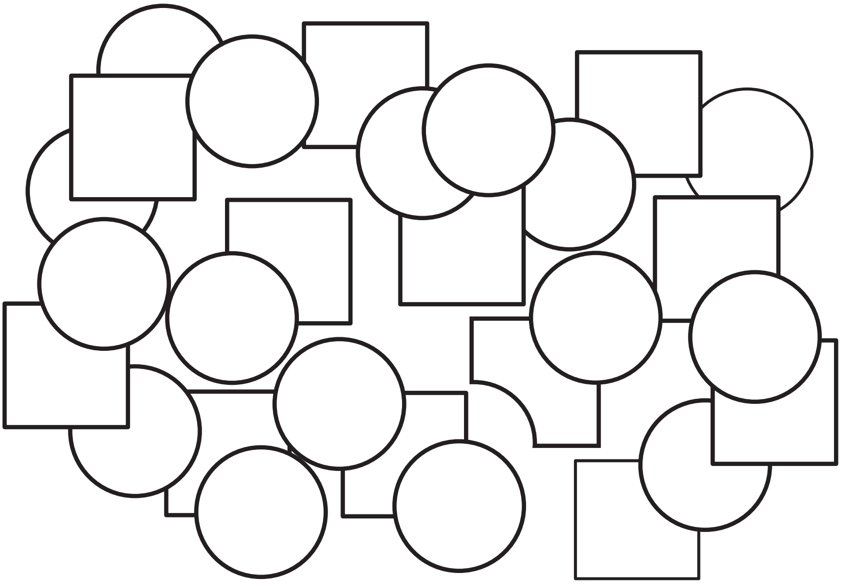


Figure 1.6
The "search set." Find two copies of the target in this set. Have the figures been "completed" in some inner display, making the target that appears to be partly occluded harder to find?

pie-shaped figures in figure 1.5 not only has the phenomenal content of being an opaque surface in front of some disks; it also appears to be brighter than the background by objective psychophysical measures and leads to the Pogendorff illusion shown in the figure. Why would these particular properties (and a large number of other such objective properties) occur *at particular locations* in the display if not because the illusory lines and the surface they define are *actually* present somewhere in the brain and provide the locations where the effect is localized? The "somewhere" in all these examples ends up being the "mental image" or Cartesian display.

The idea that the mind gets to look at a display that has been filled in and built up from separate segments is widespread. Not only is such a display thought to cover a larger spatial extent than the fovea, but it also appears to involve visual information that is no longer present on the retina, though it may have been present in the recent past. In an interesting and thoughtful essay, Julian Hochberg (1968) makes a case that many principles of visual organization seem to hold over arrays larger than those on the retina. He speaks rather cautiously and hesitantly of a visual "immediate memory," though making it clear that such a storage does not retain information in a strictly *visual* or *image* format. One reason why Hochberg speaks of a visual memory at all is that visual forms can be discerned when there are no literal forms on the retina—at least not in the sense of contours defined by luminance gradients—and so it is natural to assume that they must be in some postretinal storage. Here are some examples. In the following, I use the neutral term "discern" instead of "perceive," since I don't want to prejudge whether these count as bona fide cases of visual perception.

Forms can be displayed as contours, dotted lines, or in some cases just the high-information regions, such as vertices alone (figure 1.7).

Forms can be discerned in a field of elements if the subset of elements that lie on a particular (virtual) contour are readily distinguishable—say if they are a different shape or brightness or color from the other elements, or if they are briefly displaced (or wiggled) back and forth. Once again, in this case the form is perceived providing only that the differences are sufficient to constitute what are called "popout" or automatically registered differences (more on this in chapter 5). Figure 1.8 is an example.

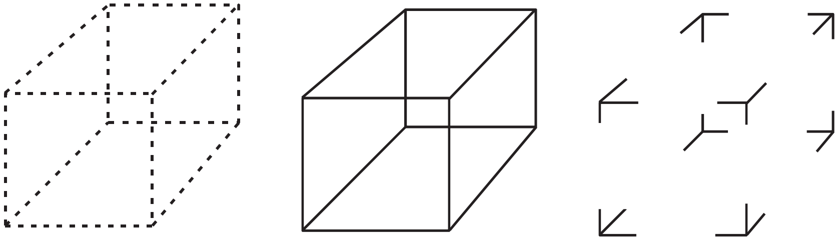


Figure 1.7
Different ways to show a Necker Cube, all of which exhibit equivalent information and lead to similar visual effects.

Forms can also be discerned in random-dot stereograms—an interesting form of visual display invented by Bela Julesz (1971). In these binocularly viewed displays, the perception of a form derives from the retinal disparity of certain regions of the display. Neither eye receives form information, but the distribution of random dots on the two eyes is such that when most of the points on the two retinas are matched, the location of the remaining points in a certain region are discrepant by some retinal distance. This discrepancy (known as "retinal disparity") is what produces the effect of stereo depth in normal binocular vision. When the region of discrepancy is chosen to correspond to a contour region, such as one that defines the line drawing of a Necker cube, a cube is perceived.

Forms can even be discerned if an opaque screen with a narrow slit in it is moved back and forth over a stimulus in a device known as an "anorthoscope." If the motion is fast enough, it appears to "paint" an image that can be discerned, though whether it is actually "perceived" remains an open question. It is reportedly even possible to recognize a form if the stimulus is moved back and forth behind a screen, so that the form is viewed as a stream of views all occurring at a single vertical line on the retina (although the phenomenal impression is not nearly as clear). Perception with these sorts of presentations has been referred to as the "eye-of-the-needle" or the "Zollner-Parks" phenomenon (see figures 1.15 and 1.16 for illustrations of forms presented through an anorthoscope).

The same form can be discerned, more or less clearly and vividly, in all these cases despite the enormous differences in the physical stimuli

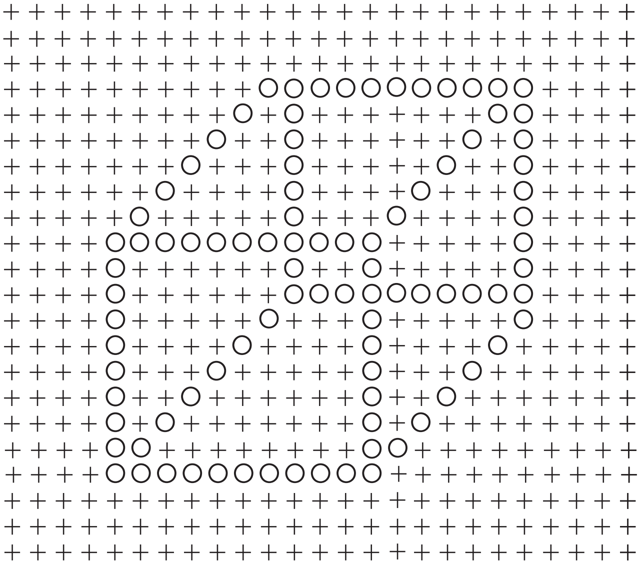


Figure 1.8
The same shape as shown in figure 1.7, but created by local feature differences. Can you *see* the form? Does it look three-dimensional, and does it reverse as you watch it?

and despite the fact that in some of the presentations, such as the random-dot stereograms and the anorthoscopic presentation, *no form at all is present on the retina*. What is important, however, is not just whether the form is recognized, but whether it exhibits the properties associated with what we call early (automatic) vision. As we will see, some of these modes of presentation do, whereas others do not, depending on how quickly they are presented and whether they are distributed over a fixed space. These signature properties of vision include the familiar 2D Gestalt grouping principles (such as grouping by proximity, similarity, common fate, and so on), as well as such 3D principles as apparent motion in 3D and the automatic interpretation of certain contours, shading, and motion cues as depicting a 3D form. The signature properties also include "perceptual coupling" between how parts are interpreted or labeled (see the discussion of labeling constraints in chapter 3). Because of this coupling of interpretations, when the interpretation of one part of a form changes, one can get a spontaneous global change or reversal of a percept, as in the Necker cube (this reversal is experienced in both figure 1.7 and figure 1.8). When the figure spontaneously reverses, the interpretation of individual edges changes, as well as the relative size of the faces, which depends on which face is perceived as the front face and which as the rear face. Other properties of early vision are discussed in chapter 2 (especially pp. 66–68).

Although, as we will see in the next chapter, many of the details of these claims are problematic in important ways, the basic idea appears sound enough: various principles of form perception and of visual organization seem to apply to a unit of display that goes beyond the current instantaneous content of the retina, and so must necessarily include visual memory. This provides some reason to think that visual processes apply to the contents of something like a "visual store," which is precisely the inner display I have been arguing against. What these examples do not demonstrate, however, and what I shall argue is not true, is that the information in the visual store is pictorial in any sense; i.e., the stored information does not act as though it is a stable and reconstructed extension of the retina.

Other reasons for postulating an inner display are sometimes given as well. For example, Kosslyn (1994) justifies his postulation of an inner screen (which he later uses to develop a theory of mental imagery in which images are projected onto this screen) by arguing that such a display is independently needed to account for visual stability and for the ability to recognize objects regardless of their location on the retina or their retinal size.[2] According to this argument, if you have a central display, you can expand or contract patterns or move them around (or, equivalently, move an "attentional window" around) so that they can be brought into correspondence with a template in a standard location on the inner screen, even if you can't do so on the retina.

[2] In fact, there is evidence that even this apparently bland assumption—that we can recognize patterns irrespective of their retinal locations—may be false in some circumstances. Nazir and O'Regan (1990) showed that if the learned pattern was of a particular size and retinal location it generalized very poorly to patterns of different sizes and retinal locations. Also Schlingensiepen, et al. (1986) showed that even simple patterns could not be distinguished without eye movements so that a static retinal location is a hindrance to pattern perception. Learning of other perceptual phenomena, such as stereopsis, generalizes very poorly to new retinal locations (Ramachandran, 1976) and retinal orientation (Ramachandran and Braddick, 1973).

But as we have seen, there are plenty of reasons to reject the idea of a central display as a way of fusing partial and fleeting images into a coherent large-compass percept. Most vision scientists do not talk about an inner display, and may even be embarrassed when confronted with the fact that their way of talking about certain phenomena appears to tacitly assume such a display. A few, like Kosslyn, actually do explicitly endorse an "inner picture" assumption. Kosslyn (1994) provides a clear case of someone who has built an elaborate theory around the assumption of an inner screen. As he puts it, "If certain properties of the world are internalized, are embodied by properties of our brains, many problems may be solved relatively easily" (1994, p. 85). This assumption will be brought under scrutiny in various places in this book. Later in chapters 6 and 7, I will examine the plausibility of a theory of mental imagery that posits the projection of images onto an inner screen. For the time being, I wish simply to look at the reasons why vision itself does not require such a postulate, and indeed why the theory ought to shun it despite its intuitive plausibility.

1.4 Problems with the Inner-Display Assumption: Part 1, What's in the Display?

1.4.1 How is the master image built up from glances?

We have seen that a form is perceived even when the retinal image is highly impoverished (and perhaps even nonexistent) and even when it is known that the retinal information is not being communicated to the brain (as in the case of the blind-spot or off-foveal parts of the display). We have also seen that off-retinal information combines in some ways with the retinally present information to produce a characteristic percept. All this suggests that information processed by the visual system comes not only from the retina (or the fovea) but also from some form of visual storage. But how does the information get into the storage? For years the common view has been that a large-scope inner image is built up by superimposing information from individual glances at the appropriate coordinates of the master image: as the eye moves over a scene, the information on the retina is transmitted to the perceptual system, which then projects it onto an inner screen in the appropriate location, thus painting the larger scene for the

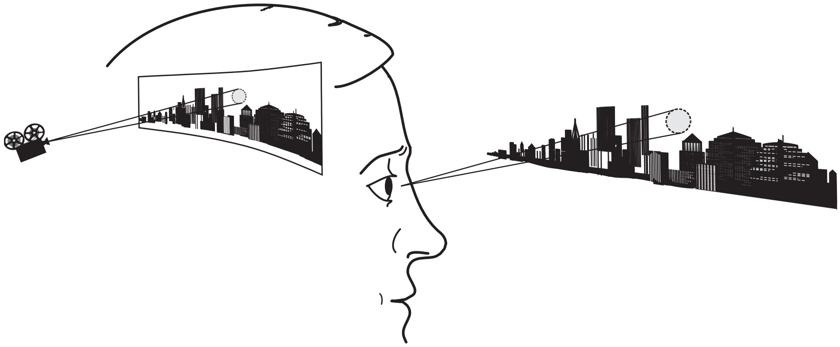


Figure 1.9
The "inner display" explanation of why we appear to see a large panorama, despite the fact that the information the brain receives is limited to a small region at the center of the field of view that is constantly moving across the scene. The idea is that an inner projector moves in registration with the motion of the eye and creates a large and detailed inner image of the scene. This intuitively appealing idea has now been discredited.

"mind's eye" to observe. The general idea behind this view is illustrated in figure 1.9.

This sort of mechanism would clearly explain both the apparent completeness and stability of the percept. This view even had some support from neurophysiological evidence showing that the locus of various visual responses in the brain (the receptive field of visual neurons) shifts when the eye is moved. This theory also received support from the widely accepted "corollary discharge" theory, which claims that when the eye is commanded to move, a copy of the eye-movement command (called the "efference copy") is sent to the "inner projector" and determines where the new information is to be overlaid (an idea that goes back to von Holst and Mittelstaedt, 1971/1950). It has been claimed, for example, that when one tries unsuccessfully to move one's eyes (when, for example, the eye muscles are injured and unable to carry out the command to move), the world appears to move in the opposite direction, since the efference copy of the command tells the projector to place the perceptual signal from the eye where the eye would have been looking had it worked properly. It should be noted here that there is much wrong with this story, not the least of which is that there is serious doubt that the position of an object appears to move to the left when the eye is commanded to move to the right but is unable to. It appears that this widely cited phenomenon may be false—as the amazing experiment by John Stevens and his colleagues (1976) seems to show. Stevens had himself totally paralyzed with curare (except for part of his arm, through which he was able to signal his replies—or call for help!) and performed the experiment in an iron lung. He reported no reverse motion of his percept when he attempted to move his eyes.

More recently, all aspects of this inner-display view have run into serious difficulties, and now the notion of superposition appears to be totally untenable. There are a number of reasons for the demise of this view of how the stable master image is built up.

Recent studies using eye-tracking equipment have provided some rather surprising findings regarding the amount of information taken in at each glance. Carlson-Radvansky (1999), Grimes (1996), Irwin (1991, 1993, 1996), and McConkie and Currie (1996) have shown that very little information is retained from one glance to another when the eyes move, or even when the eyes do not move but the display disappears briefly (Rensink, 2000; Simons and Levin, 1997). If the scene being viewed is changed in even major ways during a saccade, the change goes unnoticed. Observers do not notice changes in the color or location of major parts of a scene (unless they were explicitly attempting to examine those parts), nor do such changes have any consequence on what is perceived. Irwin (1996) showed that very little qualitative information is retained about a simple pattern of dots from one glance to another, and the location of only about 4 or 5 salient points is retained.[3]

A sequence of retinal images does not appear to be superimposed. Experiments have been carried out (O'Regan and Lévy-Schoen, 1983) in which different patterns were presented at known retinal locations before and after a saccade. What observers saw in these cases was not the superposition of the two patterns, as would be expected from, say, the presentation of the figures shown in figure 1.10 when there is a saccade between the two parts of the displays.

[3] Recent evidence suggests that accurate information tends to be available from places close to where the eye fell during recent fixations while scanning (Henderson and Hollingworth, 1999). Nonetheless the fact remains that what is retained in immediate memory is generally far from being the sort of detailed pictorial information required by the picture-painting or superposition view.

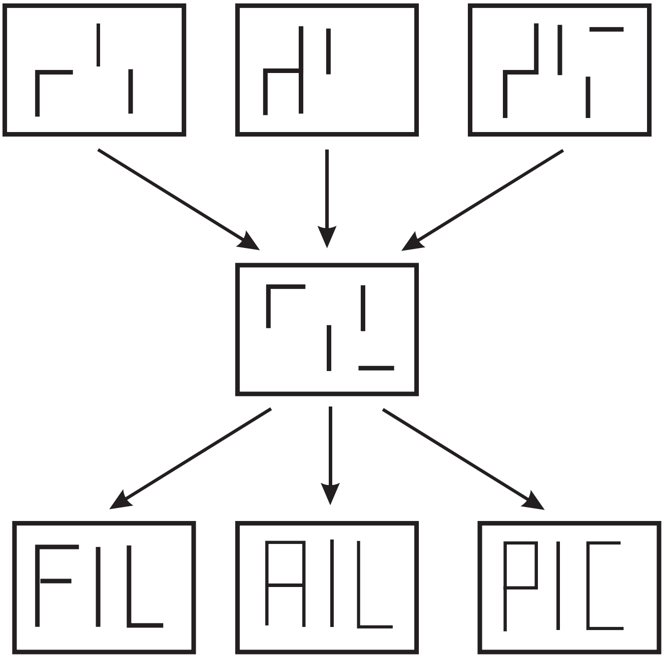


Figure 1.10
In this study described in O'Regan and Lévy-Schoen (1983), an eye movement occurs between presentation of the top figure and presentation of the middle figure. If the two were superimposed, one of the three bottom ones would be seen. There was no evidence of such superposition. (Adapted from O'Regan and Lévy-Schoen, 1983.)

1.4.2 What is the form of nonretinal information?

Despite Hochberg's observation that off-retinal (stored) visual information shows some of the principles of perceptual organization, many important visual properties are not observed when the critical interacting parts are not simultaneously in view, and even those that are observed do not have the phenomenal clarity that they have when they are actually viewed retinally, raising the question of whether they are seen or inferred (see the next section). For example, many of the signature properties of visual perception—such as the spontaneous interpretation of certain line drawings as depicting 3D objects, spontaneous reversals, recognition of the oddity of "impossible objects" such as those in Escher drawings or the so-called Devil's Pitchfork (figure 1.11) do not occur if

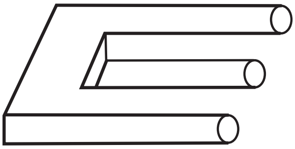


Figure 1.11
In the figure above, it takes just a brief inspection to see that something is amiss—this "devil's pitchfork" cannot be given a consistent 3D interpretation because local information leads to several of the edges having a different interpretation at each end—in other words, the edges receive incompatible labels from local interpretations.

the drawings are made large enough so that the ends are not simultaneously present on the retina. Thus, for example, such well-known figures as the Necker Cube do not appear as reversing 3D shapes and the Devil's Pitchfork does not seem so odd when it is drawn elongated and viewed in such a way that the ends are not simultaneously in the fovea (figure 1.12). Since the phenomenal percept in these cases, as in all perceptual experience involving eye movements, arguably does cover the entire object,[4] the entire object is presumably displayed on the inner screen or the master image.

Other evidence for the claim that off-retinal information (or perhaps I should say "off-foveal information") does not function in the same way as foveal information was obtained by Peterson and Gibson (1991) and Peterson and Hochberg (1983). Using figures such as those in figure 1.13,

[4] The question of what is contained in the "phenomenal image" is problematic, to say the least. I am using the term the way many theorists do, although some careful observers, like Hochberg, make a point of emphasizing that the phenomenal experience of what is sensed is quite different from information in memory. Thus, in describing what he saw through the anorthoscope aperture view (such as shown in figure 1.15), Hochberg says:

> Let me describe what our . . . aperture view looks like to me: In the aperture itself [is] a clearly *sensory* vertical ribbon of dots . . . ; the ribbon of dots—still quite clear—is part of an entire (largely *unseen*) surface of dots that is moving back and forth *behind* the aperture. . . . There is no real sensory quality to either the shape or its background, where these are occluded by the mask. I'm completely certain that I only *see* those portions of the shape that are behind the aperture at any moment, but I'm equally certain of the extension of the shape behind the mask. Is this "perception," "apprehension," "imagination"? Perhaps we're not dealing with perception at all, in these situations. Maybe merely *knowing* what the pattern is, is sufficient to elicit the different tridimensional ratings, regardless of how this knowledge is gained. (1968, pp. 315–316)

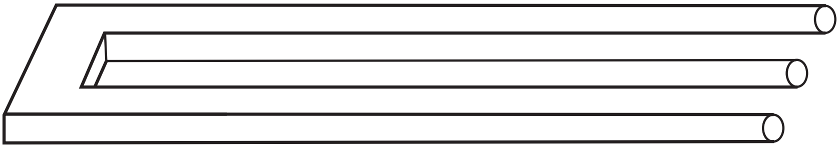


Figure 1.12
In this version, if the picture is held close up so the two ends are not simultaneously on the fovea, it is not nearly so obvious that something is wrong. Integrating the information from the two ends requires an appeal to memory; it is not just a matter of "painting" the larger picture onto the master image.

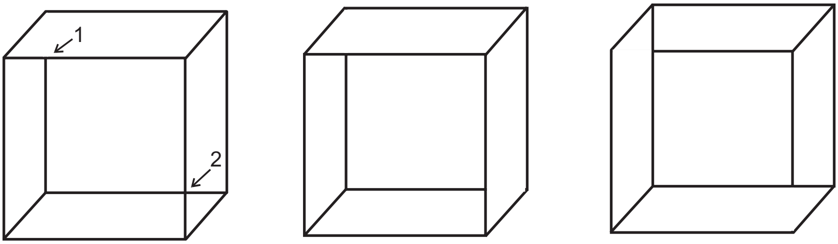


Figure 1.13
The figure on the left is globally unambiguous, yet when attending to the point marked "2" it remains ambiguous between the two orientations shown in the middle and right figures. (Adapted from Peterson and Hochberg, 1993.)

these investigators showed that the figure maintains its ambiguous status and exhibits reversals even though a part of the figure is unambiguous and therefore the entire figure would be unambiguous if the relevant cue that disambiguates the figure were taken into account. In the figure on the left, point 1 disambiguates the figure so that its shape must be that depicted by the middle figure. Yet when attending to point 2, the viewer sees the orientation of the figure alternate between the version shown in the middle and the one shown on the right.

In this example, point 1 should be able to disambiguate the entire figure, since it makes the local portion of the figure univocal. Yet it does not appear to affect how the figure as a whole is perceived; if you focus at point 2, the figure remains ambiguous. In fact, if the distance between the cue and the ambiguous parts is great enough, it has little effect in disambiguating the percept, as can be seen if we elongate the globally unambiguous figure (see figure 1.14).

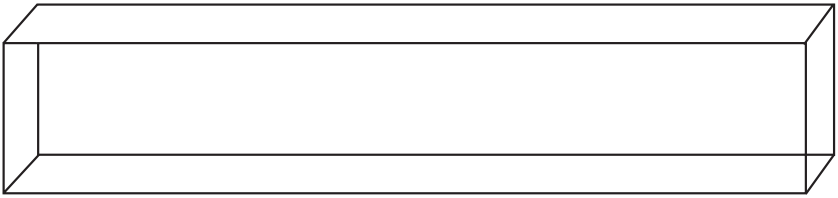


Figure 1.14
This figure, in which the disambiguating cue is farther away from the locally ambiguous parts, is even less likely to be perceived as the unambiguous box such as the middle figure of figure 1.13.

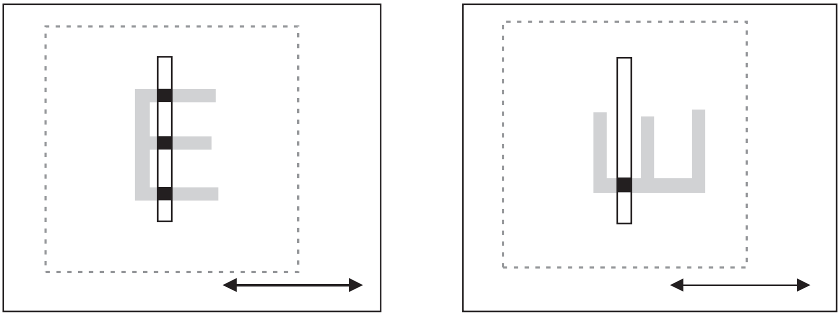


Figure 1.15
In the anorthoscope effect, a slit moves back and forth in front of the pattern. If the speed of the slit is just right, the pattern is perceived. However, the pattern is more easily perceived when there are fewer segments that have to be tracked as they pass in front of the slit (recognizing the **E** requires keeping track of three segments and checking whether they are joined, whereas recognizing the rotated **E** only requires tracking one segment and counting the number of crossbars that are passed). (This example is due to Ian Howard and is used with the author's permission.)

The same point can be illustrated by presenting visual patterns in rapid sequence to the eye. As I already remarked, in such cases observers typically feel that they see some larger integrated pattern. I have already mentioned the anorthoscope and the Zollner-Parks phenomenon, studied extensively by Parks (1965) and Rock (1981). In these studies, a pattern, viewed through a moving narrow slit that travels back and forth across the pattern, appears to be seen if the slit moves sufficiently rapidly (see figures 1.15 and 1.16). In fact, it has even been reported as perceived, though not quite as readily or clearly, if the slit is held fixed and the pattern is moved back and forth behind it. Of course, the moving-slit

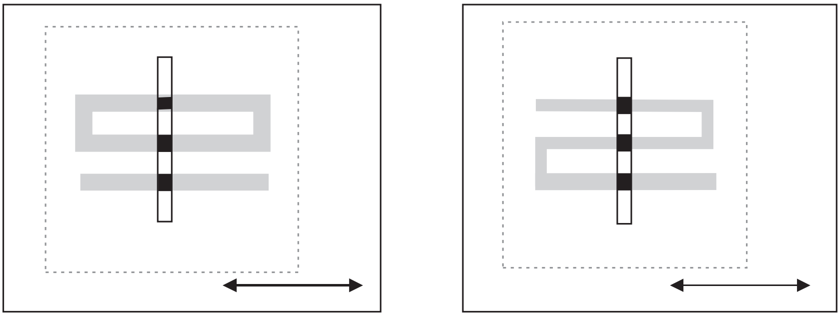


Figure 1.16
Another anorthoscope task. Observers are asked how many distinct line segments there are. These two displays have exactly the same inventory of local features (right angles, vertical and horizontal line segments, etc.). Recognizing which figure has one continuous line segment and which has two separate segments requires keeping track of which currently visible segment was connected to which segment in an earlier part of the viewing sequence. (Example due to Ian Howard.)

version could involve something like a persistent "retinal painting" by the moving-slit display, just as painting a scene on a TV set results in a display larger than the moving dot, though this is unlikely in the fixed-slit, moving-display version of the experiment. Some studies have controlled for the eye movements that would be required to paint the figure across the retina. Also, Rock (1983) showed that "retinal painting" is not in itself a general phenomenon, since simply moving a point of light along a path identical to the one that was traced out in the anorthoscope experiment does not yield a perception of the form. It turns out that the slit itself must be visible in order to get the anorthoscope effect. Not only must the slit be seen in outline; it must also be seen to be occluding the figure as the screen moves over the figure. If the visible portions of the form (the little bits that can be seen through the slits in figure 1.15 and figure 1.16) do not extend to the very edge of the slit, the effect is not observed (as illustrated in figure 3.11, to be discussed later).

Since the anorthoscope effect does not appear to be due to retinal painting, the natural assumption is that the pattern is instead being painted on an inner image, using some unspecified cues as to the motion of the figure behind the slit. But there are many reasons to reject such a view. One is that the ability to see the pattern in the case where the pattern is moving and the slit is fixed depends very much on the memory load imposed by the task of tracking the pattern. For example, in a series of unpublished studies, Ian Howard showed that patterns in which fewer features had to be tracked as they moved across the slit were identified more readily than ones that required more features to be tracked, even when the pattern was actually the same. Thus, for example, in figure 1.15, an E was harder to identify than the same shape lying on its back: the former requires that three segments be tracked as they move behind the slit, while the latter requires only one (together with a count of how many verticals went by). So the image is not just being "painted" on a master image, but must be remembered in a way that is sensitive to how many items there are to recall.

Figure 1.16 shows more clearly what must be remembered as the shape is moved past the slit. In this example, the task is to say whether there are one or two separate curves in the partially seen shape. Clearly, what an observer must do is keep track of the *type* of each line segment as it passes by the slit. This keeping track of line types—and not the operation of the visual system—is precisely the basis, I claim, for all of the demonstrations of "seeing" shapes through the anorthoscope. Such type tracking will be discussed in chapter 3 in terms of the notion of "label propagation." Once again, in the forms shown in figure 1.16 the task is easier when the forms are turned by 90 degrees, since fewer labeled lines must be tracked in that case.

Julian Hochberg (1968) conducted related studies involving serial presentation of patterns. He presented sequences of vertical slices of ambiguous figures, such as the Necker cube, at various speeds. There were two conditions in slicing up the image. In one case, it was sliced up so that slices contained complete vertices. In the other case, slices were made through the tips of the vertices so that the slices contained primarily straight-line segments (thus individual vertices were broken up in the process). The slices were presented at different speeds. Hochberg found that at fast speeds (around half a second to present 6 slices) the figures were perceived equally easily for both types of slices, consistent with other findings of a very-short-term visual buffer. But at slow speeds (more like natural viewing of these figures, in which the entire figure takes 2–3 seconds to examine), only the slices that kept vertices intact provided the information for the perception of tridimensionality.

Similar studies showed that the order in which parts of a figure were displayed through a stationary peephole made a difference in how diffi-

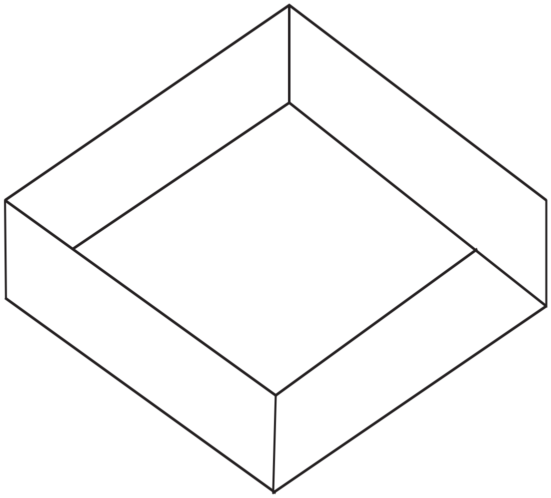


Figure 1.17
Simple "impossible figure" studied by Hochberg (1968) with the sequential presentation shown in figure 1.18.

cult it was to perceive the figure. For example, Hochberg (1968) also studied the perception of anomalous ("impossible") figures when the figure was presented in a piecemeal fashion. He found that anomalous figures (such as figure 1.17) could be detected in sequential presentations if the presentation sequence allowed observers to trace the type of the edges past ambiguous vertices until they reach a vertex where those labels are inconsistent with the requirements of a possible 3D vertex.

An example of a sequence that enables detection of the anomaly is shown in figure 1.18. In this case, however, we do not need to assume that a picture of the global pattern is being built up, because a much simpler explanation is available. It is the idea that observers are keeping track of the type of each line or edge and tracking this edge type from vertex to vertex.[5] The process is more like observers thinking to themselves, "I see this vertex as concave up, so this edge here must be an outside concave edge and must continue to be so when it gets to the next vertex. But if *that*

[5] The term "line" is generally used in reference to 2D visual features. When lines are interpreted as parts of 3D objects they are more appropriately referred to as "edges." I will try to maintain this distinction, despite the fact that whether something is a line or an edge is often unclear in many contexts. The same is true of the pair of terms "vertex" and "junction" with the former being a 2D feature.

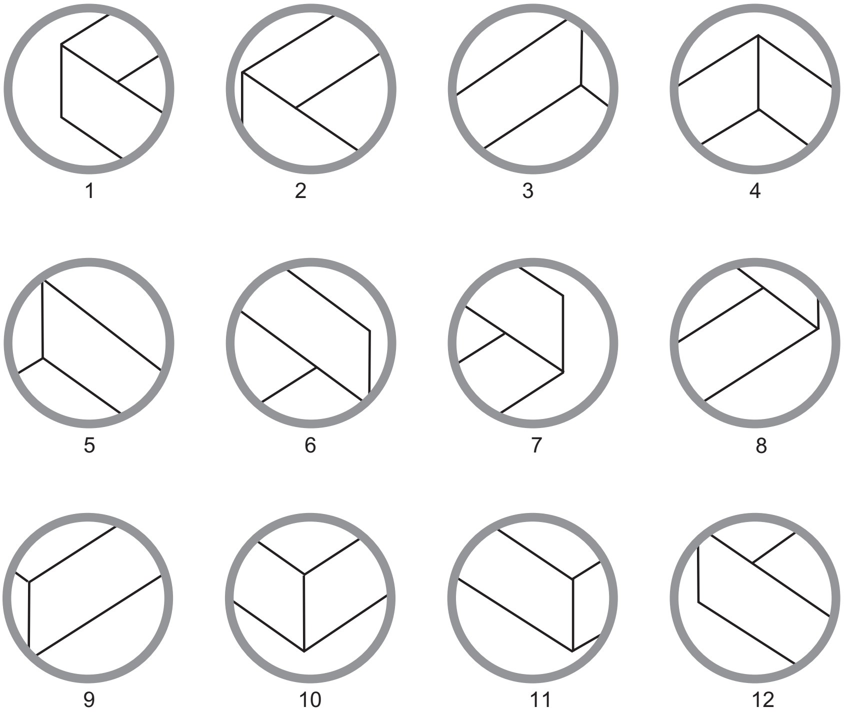


Figure 1.18
Sequence of views similar to that used by Hochberg (1968). Observers were able to detect that the sequence was from a drawing of an impossible figure only if the sequence was presented in the right order. (Adapted from Hochberg, 1968.)

edge is an outside convex edge, then *this* connecting edge must be a concave outside edge as well," and so on. In this way, if the first vertex is seen to reverse, then the rest of the labels change to maintain consistency. Note that such reasoning involves indexical (locative) terms like "this edge" and "that vertex." For such a reasoning sequence to be possible, there must be some way to refer to particular elements in the scene, and that indeed is the focus of a theory of visual indexing, to which we will return in chapter 5. For present purposes I wish merely to point out that the Hochberg experiments, like the anorthoscope examples discussed previously, all point to the importance of the notion of labeling features in a scene and tracking the labels along spatially contiguous elements (such as edges or surfaces) so as to determine whether they are consistent with labels assigned to other parts of the pattern.

The idea of tracking the labels assigned to edges helps to explain why some sequences are easier to see (or perceive as anomalous) than others. In addition, this labeling idea is in fact consistent with a body of research in computational vision that I will describe in some detail in chapter 3. In that context I will relate this labeling technique to an important question that arises concerning how a representation of a 3D world can be reconstructed from its highly incomplete and ambiguous 2D retinal projection. The technique developed in computational vision involves assigning possible sets of labels to the elements in a scene and then pruning the set by taking into account the constraints that must hold among such labels (e.g., the label assigned to an edge at one vertex must be a label possible to assign to that same edge at another vertex).

This provides an alternative way of characterizing the "signature" visual phenomena that led Hochberg to suggest that a "visual buffer" holds information in pictorial form. Such visual phenomena as spontaneous 3D interpretation, spontaneous reversals, and detection of impossible figures can be done by label propagation. This does not require that pictorial information be stored—only that there be a way to keep track of the label assigned to a *currently visible* edge as some vertices connected to that edge come into view and other vertices go out of view. In other words, so long as we can trace a particular edge and track its label continuously over time, we are in a position to interpret the lines as depicting a 3D object or to decide that no such object is possible. Interpretation of line drawings in 3D is a locally supported process, as the example of elongated figures shows (and as Hochberg has argued as well). The interpretation initially comes from cues provided by individual vertices alone. These assign (possibly ambiguous) labels to lines viewed as edges of 3-D objects, which have to be supported or rejected by connected vertices. This does not require any visual storage of off-foveal visual patterns. All it requires is that for each line segment currently in view, there be a record of what label was assigned to it by the vertex that just moved off the retina. This requires tracing, or otherwise identifying, the lines as they appear on successive retinal images. This sort of analysis works perfectly for the anorthoscope examples (the ones in figure 1.16 require an even simpler set of labels: simply keeping track of whether a particular line had ever been connected to any of the other lines in the figure).

1.4.3 How "pictorial" is information in the "visual buffer"?

As I suggested in the previous section, there is good reason for shunning the assumption that information in a "visual buffer" is pictorial. There is also considerable direct evidence that the information we extract from a scene does not have nearly the richness, geometrical completeness, and uniformity of detail that we associate with any kind of picture. In fact, as Bishop Berkeley argued, visual concepts are abstract and highly variable in their details, much as information conveyed by language is (e.g., we can describe what is in a scene in great detail while failing to mention where the things are or only vaguely describing their general shapes, as in "elongated roundish blobs"). If a master inner image were being painted, it is clear that it would have to have some very odd nonpictorial properties, such as labeled regions (what Dennett has described as a "paint-by-numbers" quality). As the examples of extended figures above suggests, once the information gets into the visual system (as opposed to still being on the retina), it no longer seems to function as visual inputs do, in terms of showing such signature properties as automatic three-dimensional interpretation and spontaneous reversals. As we will see in chapter 3, merely getting form information, such as where contours are located, into the visual system does not guarantee that it will serve to drive the usual interpretations, such as three-dimensional shape recovery. Indeed, I will show evidence that contour information provided by clearly perceptible differences in textures and colors does not always enable the visual system to see the form in 3D or in motion. So even if we want to persist in thinking of a master inner image, we will have to greatly modify our idea of what sorts of things can be painted on it—so much so that it will presently become clear that it's not an image at all.

It has been suggested that what we "see" extends beyond the boundaries of both time and space provided by the sensors in the fovea. So we assume that there is a place where the spatially extended information resides and where visual information is held for a period of time while it is integrated with what came before and what is coming in at present. Thus a central function of the "master image" is to provide a short-term visual memory. For this reason, looking at what visual information is stored over brief periods of time (seconds or minutes) may give us some insight as to what the visual system provides to the cognizing mind.[6] If we examine cases where people's visual memory is taxed, we can get an idea of how much detail and what *kinds* of details are registered there. When we do this, we find that the inner image becomes even less plausible as a vehicle for visual representation. Consider the following experimental results (discussed in Pylyshyn, 1978), which suggest that the information provided to the mind by the visual system is *abstract* and is encoded *conceptually*, perhaps in what has sometimes been called the *lingua mentis*, or language of thought.[7]

[6] It is generally accepted that the so-called iconic storage retains a complete and detailed image, though only for about a quarter of a second (Sperling, 1960). This is clearly not the storage system that is relevant to the arguments for an inner screen since our phenomenal visual experience, as well as the sorts of empirical phenomena discussed by Hochberg, apply over a much longer period of time and over a wider region than the retina. Some studies (Posner, 1978) have shown that during the first second or so information is transformed from an iconic to a more abstract (categorical) form (for example, it takes longer to judge that "a" and "A" are the same letter when they are presented in rapid succession, compared to when the first letter is presented a few hundred milliseconds earlier, whereas the time it takes to judge that they are typographically the same is less when the two are presented closer in time). Since it is this latter stage of storage that is relevant to our present discussion, this supports what I have been arguing, namely, that information in the visual store is abstract and categorical (e.g., it consists of labels).

[7] In a broad defense of the pictorial view (specifically as it pertains to mental imagery), Tye (1991) has criticized these examples on the grounds that (a) they only implicate memory and not the pictorial display itself, and (b) pictures, too, can be noncommittal and abstract. The first of these is irrelevant since one of the ideas I am questioning is precisely the pictorial view of memory representation. Although many proponents of the picture view of mental imagery may have given up on the assumption that long-term memory is pictorial, not everyone has, and certainly at the time of my critique such a view was widespread (see the quotations in Pylyshyn, 1973). As for the notion that images can be noncommittal and have an abstract character, this is simply a play on words. The way pictures get to have noncommittal content is by appealing to conventions by which they may be "read" like linguistic symbols. Sure, you can have a picture of a tilted beaker (such as in figure 1.19) that shows a fluid but is noncommittal about the orientation of the surface: you can paint a blur or a squiggle instead of showing the surface of the fluid, and then you can say that this information is indeterminate. But the blurring is simply an invitation *not to pay attention* to the part of the figure depicting the surface. It's like mumbling when you come to the part of the argument you are not sure about, which, come to think of it, is exactly what is going on in this proposal. In chapters 6 and 7 we will return to the popular shell game in which various ad hoc properties are attributed to the image to hide the fact that the work is being done not by the picture but by the "mind's eye" and the brain behind it.

The first of these examples comes from observing children, who are generally thought to have excellent visual memories. The reason that I present examples taken from observations of children is that we are especially interested in certain kinds of "errors" made in generalizing one visual situation to another (usually a pictorial one), and children tend to be less sophisticated about picturing conventions and so make more errors. We are interested in errors because these tell us what patterns the visual system finds to be most alike, and this in turn tells us something about how the visual patterns are represented in visual memory. First I will present the examples as illustrated in figures 1.19 to 1.24. After describing the results, I will then discuss the moral that might be drawn from them.

In a typical Piagetian task (see Piaget and Inhelder, 1957), a child is shown a tilted glass test tube containing colored water and asked to draw it, or to pick out the drawing that most looks like what she saw. In these experiments the child is most likely to select a drawing in which the water level is either perpendicular or parallel to the sides of the tube, as show in figure 1.19. (Later I will show that adults are not much better at this water-level task!)

If a child is shown a solid block, say a cube, and asked to draw it or to select a drawing that most looks like it from a set of alternatives, the child frequently chooses drawings such as those shown in figure 1.20, rather than the more conventional isometric or perspective projections (such as the Necker Cube show in figure 1.7). This idea was first described by Winston (1974) and led to experiments reported in an M.Sc. thesis by Ed Weinstein (1974).

It is a common observation that a child will frequently reverse a letter of the alphabet and draw its mirror image, as show in figure 1.21. This phenomenon is quite ubiquitous. When presented with any shape and

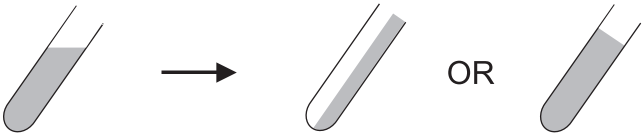


Figure 1.19
Sketch of Piaget's finding (described in Piaget and Inhelder, 1957). A child is shown the figure on the left and mistakenly recalls one of the figures on the right.

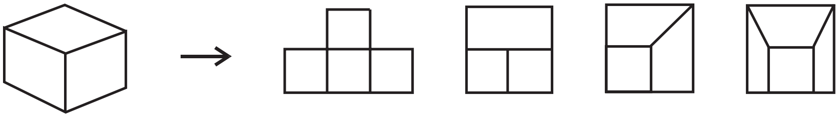


Figure 1.20
Another example of children's recall. A child is shown a real three-dimensional cube and draws one of the drawings shown on the right. (Adapted from Weinstein, 1974.)

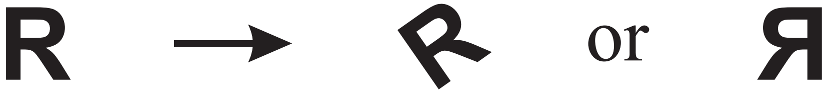


Figure 1.21
Children much more often mistake a figure for its mirror image than for a rotated version of that figure, resulting in the common reversal that has become enshrined in the name of the toy store "Toys Я Us" (Rock, 1973).

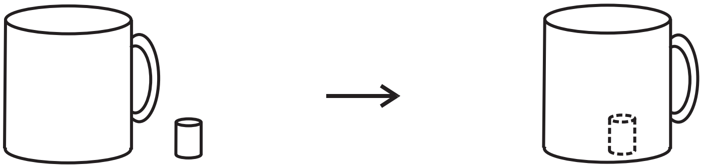


Figure 1.22
Children sometimes make what seem to us like odd errors when they imitate an adult's actions. Here a child is asked to imitate an adult placing a small object beside a cup, but in doing so places the object inside the cup (Clark, 1973). What does this tell us about how the child represents the adult's action?

asked to find the same shape among a set of alternatives, a child tends to mistake the shape and its mirror image more often than the shape and a tilted version of the shape. (Adults tend to make this error as well, though not as frequently.) These and related studies are reported in Rock, 1973.

When asked to imitate an action such as placing a small object close to a container like a cup, children more often place the object *inside* the cup rather than beside it, as illustrated schematically in figure 1.22. Imitating actions is an interesting way of examining how people (or animals) view the action. No act of imitation is ever an exact replica of the action being imitated. Not only are we incapable of perfect imitation of all muscles and movements, an imitation does not need to be a precise physical duplicate to qualify as an accurate imitation. What is required is that the imitation preserve what is essential in the action being imitated, and that in turn tells us something about how the action was perceived or encoded. These studies are part of a series reported in Clark, 1973.

The other examples are drawn from studies with adult subjects, but they illustrate the same general point. Figure 1.23 shows the results of a study on visual memory for chess positions by Chase and Simon (1973). The graph illustrates that when chess masters and novices are shown a midgame chess board for about 5 seconds, the chess master can reproduce it with almost perfect accuracy, while the novice can get only one or two chess positions correct. But when they are shown the same chess pieces arranged in a random pattern, the two groups do equally poorly. The visual-memory superiority of the chess masters is specific to real chess positions.

Figure 1.24 shows an experiment by Steve Palmer (1977) in which subjects are asked to examine two simple line drawings and superimpose them in their mind (presumably on their master images), then to select the drawing most like the superimposed combined image. There is a great

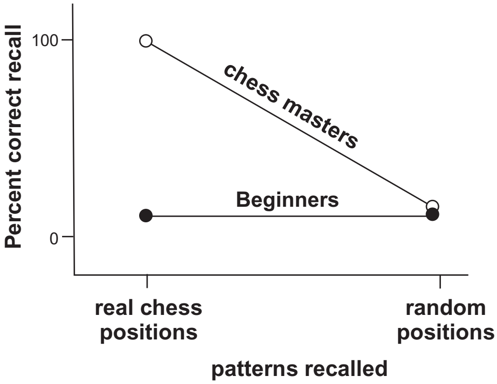


Figure 1.23
This graph shows that chess masters' apparent superior memory of chess positions occurs only when the chess pieces are arranged in a pattern taken from a real chess game. When the same chess pieces are arranged in a random pattern, the chess masters are no better than novices. (This finding is described in Chase and Simon, 1973.)

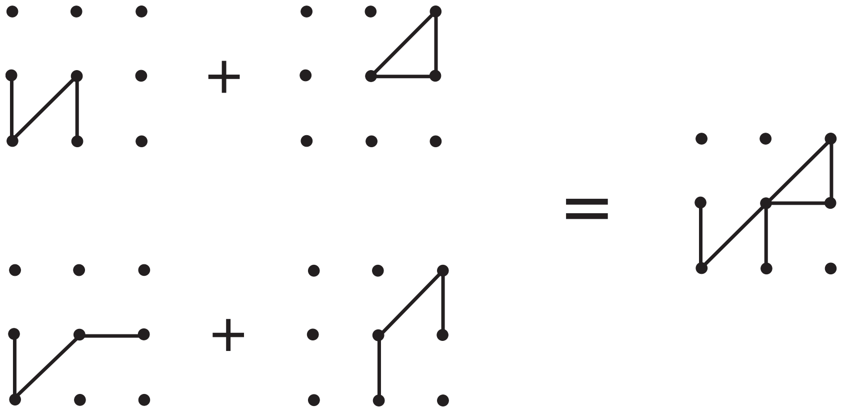


Figure 1.24
When observers are asked to superimpose the figures in the first column with those in the second column and tested as to what the combined figure looks like (in this case it's the one on the right), it matters a great deal whether the figures that are combined constitute a "good subpart" of the combined figure (which they do in the top row and do not in the bottom row*)*. (From Palmer, 1978.)

deal of difference in how well people do, depending on whether or not the two figures fit together as natural subparts to create the complex. It appears that superimposing even simple shapes in the master image is not a matter of mechanically overlaying them: the perceived subpart structure of the resulting figure—whether the two figures form natural groupings when superimposed—matters. Consequently, the two top line drawings in figure 1.24 are easier to combine than the bottom two to produce the same combined image, shown on the right.

This collection of experiments presents some perhaps surprising findings regarding errors in visual recognition commonly made by observers. What do they have in common, and what do they suggest about how the visual system encodes a visual stimulus? If the visual system constructed a master image that persisted and provided the basis for visual memory, then the errors one would expect would be something like the errors that a template-fitting process might produce. Patterns or shapes that differed least in terms of their geometry should be most often mistaken. But that is not what happens in visual memory, and it's not even what we would intuitively expect to happen. After all, when you miss-recall a scene, such as the appearance of the room full of people at your last party, you do not expect that what you will get wrong will be anything like a pictorial distortion—things moved a bit or shapes altered slightly. In fact, in the case of a two-dimensional picture, even a slight difference in vantage point would change the geometry radically without affecting what is represented in the image. People are much more likely to mistake a photograph of a room they had seen with one that was taken from a different point of view, than with one which contained a different person, no matter how large the geometrical difference is in the first case. Even if the image were three-dimensional, like a hologram, it would still be too sensitive to unimportant geometrical deviations in relation to meaningful ones. And it is the meaningful properties, which are often carried by very small pictorial details, that our visual system pays the greatest attention to. As a result, what you might forget in recalling the party scene is that Jones was to the left of Smith, though you might remember that they were close to each other and were talking. Your memory image, however complete and vivid it might seem to you, is also indeterminate and noncommittal in a large number of ways. You can recall that two people were having a good time without any recollection of what they were doing. And you can have what seems to you like a clear *image* of this state of affairs. It is possible to feel that one has a perfectly vivid and complete image of a situation that in fact is highly abstract and sketchy, and that is where one's phenomenal experience leads one astray. I often feel I have a vivid image of someone's face, but when asked whether the person wears glasses, I find that my image is silent on that question: it neither has nor lacks glasses, much as the blind spot neither provides information about the relevant portion of the visual field nor does it contain the information that something is missing. You might note that *sentences* (and other languagelike compositional encoding systems) have this sort of content indeterminacy, whereas pictures do not. You can say things in a language (including any language of thought) that fails to make certain commitments that any picture would have to make (e.g., your sentence can assert that *A* and *B* are beside one another while failing to say which is to the right or left).

In terms of the examples just enumerated, if children's visual experiences are represented not as pictures but as conceptual complexes of some sort (I will not speculate at this point what such a complex might be like, except to point out that it is more like a language of thought than a picture), then the availability of certain concepts could be reflected in the errors they make. There is no way to represent (i.e., describe) the tilted-test-tube display without a concept such as that of allocentric level (or parallel to the surface of the earth). If this concept is not available, then there is no way to capture the special feature that distinguishes between the three displays in figure 1.19. So the child is left with choosing a salient pattern as consistent as possible with what he or she sees, which happens to be a surface that is either parallel or perpendicular to the sides of the tube. Exactly the same can be said of the example in figure 1.20. If shapes are represented conceptually, rather than pictorially, distinguishing a shape from its mirror image requires access to the egocentric concept *left of* or *right of* (try describing a shape in such a way that it can be distinguished from its mirror image without using such terms or their cognates), and these ego-reference concepts are slow to develop compared with concepts like *up* or *down* or *sharp angle* or *perpendicular* or *circular*, and so on. I don't mean that the words "left" or "right," and so on, are not available, but that the underlying concepts that these words conventionally express are not available (although without the concepts, the words could not be learned either).

So long as we appreciate that what the visual system provides is abstract and conceptual, rather than pictorial, we will also not find the other results puzzling. To mimic a movement is not to reproduce it as depicted in an image, it is rather to generate *some* movement (perhaps one that is preferred on some other grounds) that meets the conceptual representation of the movement as it was seen, or as the visual system represented it. Thus if the child represented the action of moving the small object as an action in which the object was placed in some relevant and appropriate proximal relation to a cup, she might choose to place it inside just because she prefers inside-placing (there is certainly evidence that children like to place things inside other things). Similarly, the results of the chess-memory and superposition experiments sketched in figures 1.23 and 1.24 are baffling if one thinks of the visual system as providing a picture that serves as the form of short-term memory. But they are easily understood if one views the memory entry as being conceptual, where the concepts are either learned over the years or are part of the native machinery of visual organization (and for present purposes we need not take a stand on which of these it is). With the right

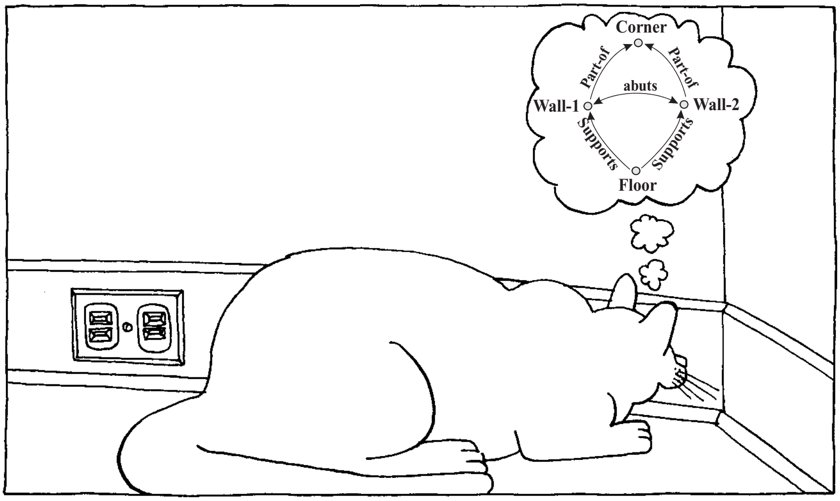


Figure 1.25
Alternative conception of what is in the mind of the cat in the cartoon in figure 1.1. B. Kliban (American, 1935–1990). From the book *Cat*, by B. Kliban. Used by permission. All rights reserved. © Judith K. Kliban.

concepts, the representation of a scene can be simple and compact and easily remembered (one may say, with George Miller, that it has been "chunked"), even if its geometrical or pictorial configuration is not.

These examples, as well as those discussed in section 1.4.2, strongly suggest that information about a visual scene is not stored in pictorial form, but rather is stored in a form more like that of a description, which is characterized by variable grain and abstractness and is based on available concepts. Thus rather than thinking of vision as it was depicted in the Kliban cartoon in figure 1.1, one should replace the picture in the thought balloon with a data structure such as that in figure 1.25, in a format that is typically used in artificial-intelligence applications.

1.5 More Problems with the Inner-Display Assumption: Part 2, Seeing or Figuring Out?

In chapter 2, I will discuss other methodologies for studying visual perception, and in particular for trying to sort out the thorny problem of which properties of visual apprehension are properties of vision as such, and which are properties of the cognitive system. I will argue that the empirical data are on the side of a clear separation between these processes, providing we are willing to make the sorts of distinctions and idealizations that are ubiquitous in science. But why do we believe in this separation, and what are we getting ourselves into if we follow this course? As I said earlier, this book is not about the nature of visual experience, as such. Yet we cannot get away without some comment on this question, because visual experience appears to be the main source of data on the operation of the visual system. Even when we appeal to the interaction of visual properties, as I did in some of the examples above, or when we use nonverbal evidence (e.g., pointing, reaching, grasping, event-related potentials, or galvanic skin responses) about perceived objects and thereby get stronger converging evidence, visual experience remains the reference point against which we measure what we mean by *seeing*. The situation is rather similar to that in linguistics, where certain signature properties of grammatical structure, such as intuitions of grammaticality and ambiguity, form the basic data, even though these have to be supplemented by theoretically motivated converging observations and judgments. In the case of vision, we must supplement our use of phenomenal experience as the data of vision, because phenomenal experience is not always available, because we don't want to tie ourselves to the assumption that only consciously experienced percepts constitute genuine vision, and because visual experience is itself a fallible source of evidence. But how can our experience be fallible: are we not the final authority as to how things seem to us? Whether or not we want to claim that we are the final authority on how things seem to us, the question of the *content of our perception* is broader that how things seem to us, because, unlike conscious experience, it is a construct that must serve in information-processing theories and must eventually comport with biological evidence.

1.5.1 A note about terminology

Many of the examples we have considered so far raise questions about when a phenomenon is truly "visual" and when it is conceptually or logically derived or based on figuring out how the world must have been in order to lead to the information we received from the senses. After all, it is possible that the reason we can recognize the words "New York" in our earlier example (figure 1.4) might simply be that we can guess them from the bits of information we can pick up; there may be nothing visual about the process and hence no need to postulate an inner completed display. In the preceding section I claimed that vision and cognition could (and should) be distinguished. But in the everyday use of the terms, the two overlap extensively. Consequently, there are those who object to using a term, such as "vision," in a way that is at variance with its general informal use. We use the term "vision" (or sometimes "early vision") to refer to the part of visual perception that is unique to vision and is not shared by cognition in general. Such a usage has been viewed by some as, at best, terminological imperialism and, at worst, circular, since it assumes that early vision is impenetrable when the very notion is defined in terms of encapsulation from cognition.

In defense of the present usage, however, it should be pointed out that it is perfectly legitimate to adopt a term that refers to that aspect of the brain's function that is distinct and uniquely associated with what goes on in a modality under study (where the exact bound of the "modality" is also an empirical issue; see pp. 126–129). To use the term "vision" to include all the organism's intellectual activity that originates with information at the eye and culminates in beliefs about the world, or even actions, is not very useful, since it runs together a lot of different processes. The same tack was adopted by Chomsky, who uses the term "language" or "language capacity" to refer to that function that is unique to linguistic processing, even though understanding natural-language utterances clearly involves most of our intellectual faculties. It is also the tack I adopted when I use the term "cognition" to refer to processes that operate over representations of knowledge, as distinct from knowledge-independent processes of the "cognitive architecture" (Pylyshyn, 1984a), when I use the term "learning" to refer to certain cognitively mediated changes in cognitive states (Pylyshyn, 1984b, pp. 266–268), and when I use "inference" to refer to any quasi-logical process.[8] Moreover, this use of "vision" is not circular (or at least not

[8] The present policy in regard to the use of the term "inference" differs from that of Fodor (1983) and Fodor and Pylyshyn (1981) in that I here do not use it to refer to processes that are systematically restricted as to what type of input they may take and the type of principles that they follow. Thus I consider early vision, which follows the sorts of principles sketched on pp. 66–68 and in

viciously so), since it embodies a strong empirical claim, namely, that there exists a nontrivial part of the overall visual process that is impenetrable. The burden of the next chapter is to argue that a significant part of the intuitive (or prescientific) sense of "visual perception" is in fact impenetrable, and that this part is also complex and covers a great deal of what is special about vision (I will discuss the question of what the visual system, so construed, outputs to other systems in chapter 3).

The reason for this terminological policy is the usual one that applies in any science. A science progresses to the extent that it identifies general empirically valid distinctions, such as between mass and weight, heat and temperature, energy and momentum, and so on. I propose a distinction between vision and cognition in order to try to carve nature at her joints, that is, to locate components of the mind/brain that have some principled boundaries or some principled constraints in their interactions with the rest of the mind. To the extent that we can factor the cognitive system into such components and can specify the nature of the interactions that are permissible among them, we will have taken a step toward understanding how the system works. For the time being, I will take for granted that showing principled macroarchitectural components can be a step toward understanding how a complex system functions (assuming that the description is valid). Given this background we can then ask, Why should we expect there to be a sustainable distinction between cognition and perception? Or more specifically, why do we think that if we draw such a boundary in a principled way, the part that is not "cognition" will include anything more than the sensors? I devote chapters 2 and 3 to arguing the case in favor of the hypothesis that vision and cognition are largely separate functions, i.e., that vision is what I and others have called a module of the mental architecture (see Fodor, 1983). This, I claim, is a major empirical discovery of vision science of the past 30 years.

chapter 3, not to merit the ascription "inference." Although it might be possible to characterize the operation of the visual system in terms of "rules," these differ significantly from rules of interence since they only apply to representations arising directly from vision and not to those with a different provenance. Because of their rigidify they are best viewed as the wired-in regularities such as any mechanism must possess. As in the case of the term "vision," something is lost by being too ecumenical in one's linguistic usage: one loses the ability to distinguish between a quasi-logical system of inferences and other sorts of causal regularities.

1.5.2 "Seeing *x*" versus "believing that what you saw is *x*"

There is an important distinction to be made between how you experience a perceptual event and what you believe about your experience, and therefore what you may report about what you saw. People's beliefs are notorious for being filtered through their tacit theories and expectations. Thus it is not clear what to make of such results as those reported by Wittreich (1959). According to Wittreich, a number of married people reported that when two people, one of whom was their spouse, walked across the well-known Ames distorted room, the stranger appeared to change in size (the usual experience), whereas the spouse did not. There are several possible explanations for this surprising phenomenon (if, indeed, it is a reliable phenomenon). One explanation (the one that Wittreich favors) is that the perception of size is affected by familiarity. Another is that a highly familiar person can result in an attentional focus so narrow that it can exclude contextual visual cues, such as those provided by the Ames room and by the accompanying person. Yet another possibility is that because of all the emotional connections one has with a spouse, it is just too hard to accept that the spouse has changed size while walking across the room. As a result, observers may simply refuse to accept that this is how their spouses appeared to them. It is not always possible to describe "how something looks" in terms that are neutral to what you know, although clearly this does happen with illusions, such as the Müller-Lyer illusion (see figure 2.3).

Consider the following related example in which a subject's report of "how things look" may well be confounded with "what I believe I saw" or "how I judge the perceived object to be." In a classical paper, Perky (1910b) reported a study in which observers were told to imagine some particular object (e.g., a piece of fruit) while looking at a blank screen. Unbeknownst to the subjects, the experimenter projected faint images on the screen. Perky found that subjects frequently mistook what they were faintly seeing for what they were imagining (e.g., they reported that the images had certain properties, like orientation or color, that were actually arbitrarily chosen properties of the faintly projected image). One way to view this is as a demonstration that when the visual experience is ambiguous or unclear, subjects' beliefs about their experience are particularly labile to alteration. In this case what the subjects sometimes believed is that they saw nothing but had the experience of imagining something. In other cases, perhaps in this same experiment, the converse obtained: subjects believed they had seen something but in fact they had seen nothing and had only imagined it. Various methodologies, such as signal detection theory, have been developed to try to drive a wedge between the factors leading an observer to decide certain things and factors leading to their detecting things with their senses (for more on the interaction of vision and "mental images," such as in the Perky effect, see section 6.5).

The point is that even if "how something looks" is determined by the visual system, what we believe we are seeing—what we report seeing—is determined by much more. What we report seeing depends not only on vision, but also on a fallible memory and on our beliefs, which in turn depend on a number of factors that psychologists have spent much time studying. For example, it is known that the larger the role played by memory, the more unreliable the report. There is a great deal of evidence that what people believe they saw is highly malleable, hence the concern about the validity of eyewitness testimony (Loftus, 1975). The often dramatic effects of subliminal stimuli, hypnotic suggestion, placebos, and mass hysteria result from the gap (things often do result from a gap!) that exists between seeing, believing, and believing what one has seen. Indeed, because of such malleability of reports of experiences, psychologists long ago came to appreciate that research methods—such as double-blind testing and the use of unobtrusive measures—had to be designed to control for the fact that honest well-meaning people tend to report what they believe they should be reporting (e.g., the "correct" answer or the answer that is wanted—the so-called experimenter-demand effect). It's not a matter of deliberately lying, although the very notion of a deliberate lie came under suspicion long ago with the recognition of unconscious motives (with Freud) and of tacit knowledge, both of which are important foundational axioms in all of the human sciences. What is at issue is not the observer's sincerity, but the plasticity of the belief-determining process. There is no sure methodology for distinguishing between what people experience in a certain perceptual situation and what they (genuinely) *believe* they experienced, although we will discuss a number of methods for refining this distinction later.

1.5.3 Reports of what something "looks like": What do they mean?

There is a further problem with some studies that build on reports of how things look and how these reports can be influenced by beliefs, utilities, expectations, and so on. A problem arises from the fact that a phrase such as "That looks like *x*" is typically used in a way that merges with something like "My visual experience has convinced me that what I am seeing is *x*." The terminology of "appearances" is extremely problematic. Wittgenstein provides a typical eye-opening example of how "looks like" runs together appearances and beliefs.

The playwright Tom Stoppard tells the story in his play *Jumpers* by having two philosophers meet. The first philosopher says, "Tell me, why do people always say it was natural for men to assume that the sun goes around the earth rather than that the earth is rotating?" The second philosopher says, "Well, obviously, because it just *looks* as if the sun is going round the earth." To this the first philosopher replies, "Well, what would it have looked like if it had looked as if the earth was rotating?"

Examples closer to our immediate concerns are easily found. For instance, it is commonly reported that how big something "looks" depends on the presence of size cues in the form of familiar objects (so, for example, when you are shown a photograph of an unfamiliar shape, it is common to include something familiar, such as a person or a hand, in the photograph). But this may well be a different sense of "looks like" than what is meant when we say that in the Müller-Lyer illusion one line looks longer than the other. In the case of the "perceived size" of an unfamiliar object, the object may not actually look different, depending on nearby size cues; it may simply be judged to be a different size.

Sometimes claims that some stimulus is "seen" in a particular way have been contested on the grounds that perception and inference have been conflated. For example, a disagreement arose between Theodore Parks and Ralph Haber regarding whether what has been called the eye-of-the-needle or anorthoscope phenomenon demonstrates "post-retinal storage" (Haber, 1968; Haber and Nathanson, 1968; Parks, 1965; Parks, 1968). In the original anorthoscope effect discussed earlier (and illustrated in figure 1.15), I claimed that people could "see" a stimulus pattern that was viewed through a slit in a screen that moved back and forth in front of the stimulus. As I already suggested, this sort of seeing is different from the usual kind in that there is a memory load imposed by the task that shows up in differences in the ability to recover the shape depending on the order in which parts of the figure are presented. Haber and Nathanson (1968) raised the question of whether what is stored in the anorthoscope effect is an image or more abstract information that allows an interpretation to be inferred (rather than seen). The question of when some episode constitutes a case of visual perception (i.e., of "seeing"), as opposed to being merely a case of drawing an inference from fragmentary visual cues, is more than a terminological one—it has implications for theories of visual memory and mental imagery.

An even more extreme case of the overly inclusive way in which the term "see" or "looks like" is used is provided by the case of "Droodles"—a type of humorous visual puzzles first developed by Roger Price, such as the ones in figure 1.26. These have sometimes been cited (e.g., Hanson, 1958) to illustrate that what you see depends on what you know. (Look at each figures and then ask yourself, What does it look like? Then do it again after reading the captions in note 9.)

Like Gestalt closure figures (or fragmented figures, discussed in chapter 2 and illustrated in figures 2.6 and 2.7), these appear to come together suddenly to make a humorous closure. But unlike the fragmented figures, these interpretations clearly depend on collateral information. The question is, Do these cases illustrate the operation of the visual system, or are they more like puns or jokes in which the punch line causes one to cognitively reinterpret or reframe what came before (or what was seen)?

Ordinary language uses terms like "appears" or "seems" in ways that do not distinguish plausible functions of the visual system from inferences based partly on visual cues and partly on other (nonvisual) information. For example, we speak of someone "looking sick" or of a painting "looking like a Rembrandt." Whatever is involved in this sort of "looking like," it is unlikely to be the basis for building a scientific theory of vision, since it clearly involves more than vision in the sense in which this term is used in science (I will return to this issue in the next chapter).

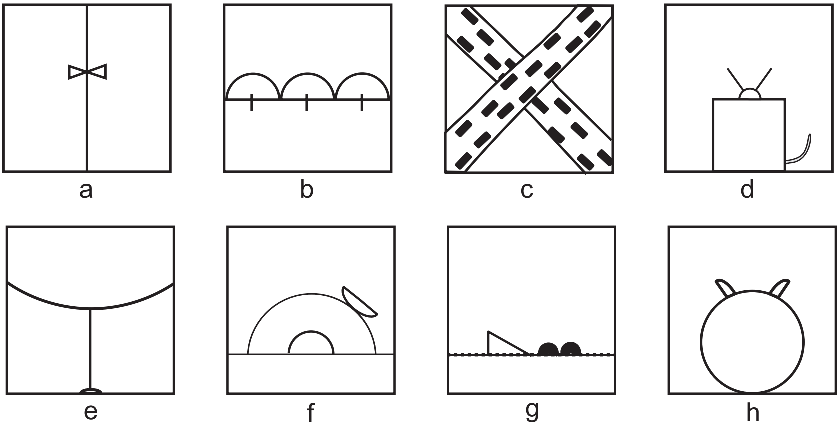


Figure 1.26
Examples of visual puns invented by Roger Price, known as "droodles." What do you see in each of these panels? Reproduced with permission from Mr. Leo Valdes, who maintains the droodles home page at http://www.droodles.com.[9]

1.5.4 Vision and conscious appearance: Can they be separated?

Although the intuitive sense of "how things look" provides the starting point in a study of visual phenomena, this is not the only way to determine what the visual system does or what it produces. For example, vision leads not only to the phenomenal experience of *seeing* (I will have more to say about this experience on pp. 350–357), it also leads to our being able to act appropriately towards objects (e.g., point to them, grasp them, and so on). When certain properties are perceived, we can also make certain judgments about the objects that have those properties; for

[9] Droodles (a), (b), and (c) are originals by Roger Price. The others are contributions (ca. 1997) to the Droodles home page maintained by Leo Valdes (http://www.droodles.com), reproduced with permission of Mr. Valdes. The original captions are:

(a) Man wearing a bow tie entered an elevator and the door closed on his tie.

(b) Rear view of the starting line for a rat race.

(c) Giraffes in love.

(d) Cat watching TV.

(e) Flea holding up an elephant.

(f) Igloo with a satellite dish.

(g) Shark returning from a holiday at Disneyland.

(h) Rabbit blowing a large bubble.

example, we can discriminate them or recognize them to be different. When surfaces with different properties, such as different colors, are placed side by side, we can judge that there is a visible boundary between them (in the extreme, when we cannot discriminate any boundary between surfaces that have different spectral properties, we say that those properties are "metamers," meaning that they are visually indiscernible). In addition, under certain circumstances we can also show that perception of properties leads to certain physiological responses (such as the galvanic skin response, which is the basis of lie-detector tests) or neurological responses (such as patterns of EEGs called event-related potentials), and so on. Another way to try to distinguish between purely visual phenomena and phenomena involving beliefs is to appeal to the interaction between two visual phenomena, one of which is independently known to occur in early vision. This is what was done in the interaction between "perceived virtual contours" and the Pogendorff illusion (figure 1.5) and when I appealed to certain "signature properties" of vision, such as automatic figure-ground separation, interpretation in 3D, reversals, and apparent motion. I also hinted at other methods, such as the use of signal-detection theory, event-related potentials, the galvanic skin response.

To understand whether certain phenomena are purely visual, we can also the appeal to clinical cases of brain damage that show deficits in reports of visual perception and in the ability to recognize objects, but without concomitant deficits in related cognitive skills. For example, there are remarkable cases of what is called "blindsight" (studied extensively by Weiskrantz, 1997), in which some patients with cortical damage have as a result large blind regions in their visual fields (often as large as half the visual field). When objects are held up before them in these "blind" regions, the patients say that they seen nothing there. Yet when they are forced to guess or to reach for the objects (just to humor the experimenter), they perform significantly above chance in both types of tasks. Such reports of not seeing accompanied by performance indicating that visual information is being processed also occur with split-brain patients (patients who had their corpus collosum surgically cut in order to alleviate epileptic seizures, or who were born without the connecting fibers). In these patients there is almost no communication between the two hemispheres of the brain, so that the left half, which has the language skills, cannot communicate with the right half, which gets input from the left half of each retina. Such people exhibit amazing symptoms (Gazzaniga, 2000). For example, they report that they do not see objects presented to the left half of their visual field. Yet their left hand (which is connected to the half of the cortex that is receiving visual information about the objects) is able to reach for the objects quite normally. In fact, they can often recognize the objects by their feel or the sound they make when moved. Once the left hand brings the object into view of the right hemisphere, these people can report seeing them. Other related visual disorders also suggest that equating seeing with being able to report a conscious visual experience may unnecessarily limit the scope of the evidence for vision.

The point is that there is no limit to the type of evidence than can in principle be marshaled to help us understand visual perception. As I already remarked, even though perceptual experience may define the clear cases, the strategy in visual science, as in all human sciences, is then to let various convergent measures and the developing body of theory determine where the boundary between perception and cognition will fall. Thus there is no reason in principle why we should not include in the category of perception cases of unconscious perception. Indeed, perhaps one might even have good reason to call certain mental states cases of unconscious perceptual *experiences*. None of these issues can be prejudged in the absence of at least a partial theory of what it is to have a conscious experience; once again, common sense is no help in these matters. The everyday notion of seeing is too fluid and all encompassing to be of scientific use. Science needs to make certain distinctions and to identify what Simon (1969) refers to as "partially decomposable" systems. But then such distinctions invariably belie the everyday prescientific ideas.

1.6 Where Do We Go from Here?

This chapter has provided a sketch of some of the reasons why many people have assumed that vision provides us with an inner version of the world more complete, detailed, and extended, and more responsive to our beliefs, desires, and expectations, than is the retinal information we are forced to deal with in the first instance. In the course of this discussion I have hinted that in formulating a scientific theory of vision, we will more than likely have to shed much of our intuitively comfortable view. In particular, we will have to jettison the phenomenal image or display and come to grips with the information-processing task that vision carries out. But we will also have to come to terms with other equally uncomfortable conceptual issues. For example, in the discussion so far I spoke freely about the *visual system* and the *visual process*. But what if there is no specifically visual process, but only an undifferentiated cognitive process. For example, many people view vision as being quite close in spirit to the process of science itself, where people use all the intellectual apparatus at their disposal to come up with theories, which they then attempt to confirm or disconfirm, leading to newer (and hopefully better) theories, and so on in an endless cycle. If this picture is correct, then it is highly unlikely that there will ever be a theory of visual perception, any more than there is a theory of science. Indeed, the nature of the scientific process, or the problem of induction, remains one of the most difficult puzzles in philosophy. But we are here embarked on a more optimistic venture: I will defend the thesis that there is such a thing as a *visual system*, apart from the entire system of reasoning and cognizing that humans and other organisms possess. I will examine the claim that vision is, as Fodor (1983) puts it, a *module*, informationally encapsulated from the rest of cognition and operating with a set of autonomously specifiable principles. More particularly, I will argue that an important part of what we normally would call visual perception is *cognitively impenetrable*. Earlier I suggested that it is problematic to distinguish between vision and visual memory because space and time can be traded off in the visual process (as happens routinely when we scan our eyes around). What I now want to claim, in contrast, is that there *is* a distinction between "seeing" and "thinking."